빅데이터 Big Data

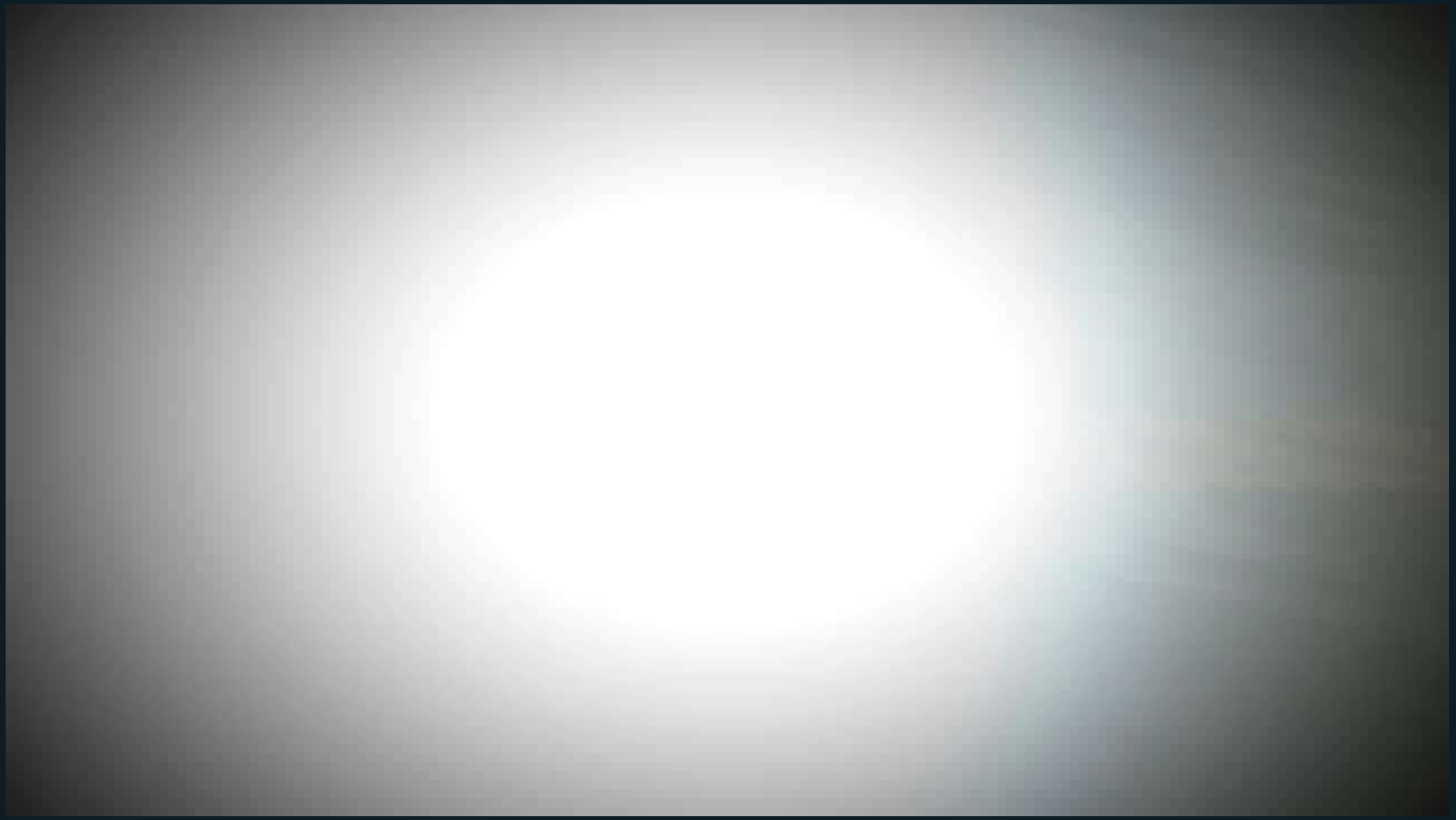
소프트웨어와 미래사회

2019

시간 속에 모든 가능성이 존재한다.
미래를 예측하는 최고의 방법은
미래를 만드는 것이다.

– Alan Curtis Kay

The best way to predict the future is to invent it



빅데이터 - 세계미래포럼 (2014)

[선택이 아닌 기본 역량 Data Literacy]

데이터 리터러시 (Data Literacy)

- 데이터를 읽고 그 안에 숨겨진 의미를 파악, 분석하여 목적에 맞게 활용할 수 있는 능력
- 다양한 분야에서 데이터를 활용하여 가치를 창출해 내기 위한 시도 증대

"산업혁명 시대에는 빅데이터가
혁신의 핵심이 될 것이다."

- 클라우스 슈밥세계경제포럼회장-

[빅데이터 (Big Data) 출현배경]

분산처리기술

소셜 네트워크

Big Data 출현
환경적 요소

스마트 기기

클라우드 컴퓨팅

[FAANG의 뒤를 잇는 MAGA]

MAGA (Microsoft, Amazon, Google, Apple)

- **2018년 하반기 들어 미국IT 산업을 이끌 던 FAANG를 대신하는 신조어**
- **페이스북 주가가 하락하자 FAANG에 대한 우려가 커지면서 새롭게 주목**
- **매출이 여러사업에서 골고루 발생하고, 미래를 위한 신성장 동력을 확보**
- **클라우드(Cloud)에 집중하는 전략을 가지는 공통점**
  - Netflix는 가입자 증가세 둔화로 주가하락 압력을 받기도 하나
    MS는 지난 1년동안 Netflix 시가 총액의 2배수준인 2천 800억 달러 증가

Microsoft amazon Google Apple

[전세계에 구축하는 데이터 센터]

- 전세계에 데이터센터를 구축한 구글
- 두번째로 큰 오클라호마 센터가 두번째로 큰 곳
- 축구장 일곱개 넓이. 3중보안
- 풍력으로 전기 생산하고, 열을 식힘

- 강원도 춘천에 위치한 네이버 데이터센터
- 초당 약 7,400개의 검색어 저장,
  약 470건의 이미지 처리




- 자연 냉각이 가능한 페이스북 데이터 센터
- 북극과 70km정도 떨어진 스웨덴에 설치

- 해저 수중 데이터센터 구축 프로젝트(Natick) 진행
- 조류를 이용한 전력변환과 해저 낮은온도를 이용한 냉각

BIG DATA

[데이터와 단위]

- 수십 테라바이트(Tera byte) 혹은 수십 페타바이트(Peta byte) 이상이 빅데이터의 범위에 해당
- 1 TB는 1024 GB, 1 PB는 1024 TB
- 1 PB는 6 GB DVD 영화를 17만 4천 편 담을 수 있는 용량

Killo-Mega-Giga-Tera-Peta-Exa-Zeta-Yotta

[빅데이터 (Big Data)]

빅데이터의 공통적 속성 3V

- 다양한(Variety) 형태로 수집&저장된
- 대용량(Volume)의 데이터들을
- 빠른 속도(Velocity)로 분석

- 데이터의 일정한 패턴을 찾아내고,
- 새로운 가치(Value)를 창출하는 것.

구분	기존 데이터	빅데이터
데이터 양	테라바이트(TB) 수준	테라바이트(TB)~제타바이트(ZB) 수준
데이터 유형	정형 데이터 위주	정형 데이터 및 비정형 데이터 모두 포함(비정형 데이터의 비중이 높음)
처리 과정	· 처리 과정이 단순함 · 원인과 결과 관계를 규명하는 데 중점을 둠	· 처리 과정이 복잡하고 분산 처리 기술이 필요함 · 상관관계를 규명하는 데 중점을 둠

[출처] 빅데이터 동향 및 정책 시사점, 제25권 10호 통권 555호, 2013. 6.

[빅데이터 (Big Data)의 속성]

빅테이터의 새로운 V

- **정확성(Varacity)**

  데이터의 신뢰성.

  데이터가 많아질수록 엉터리 데이터도 증가할 가능성이 높음

- **가변성(Variability)**

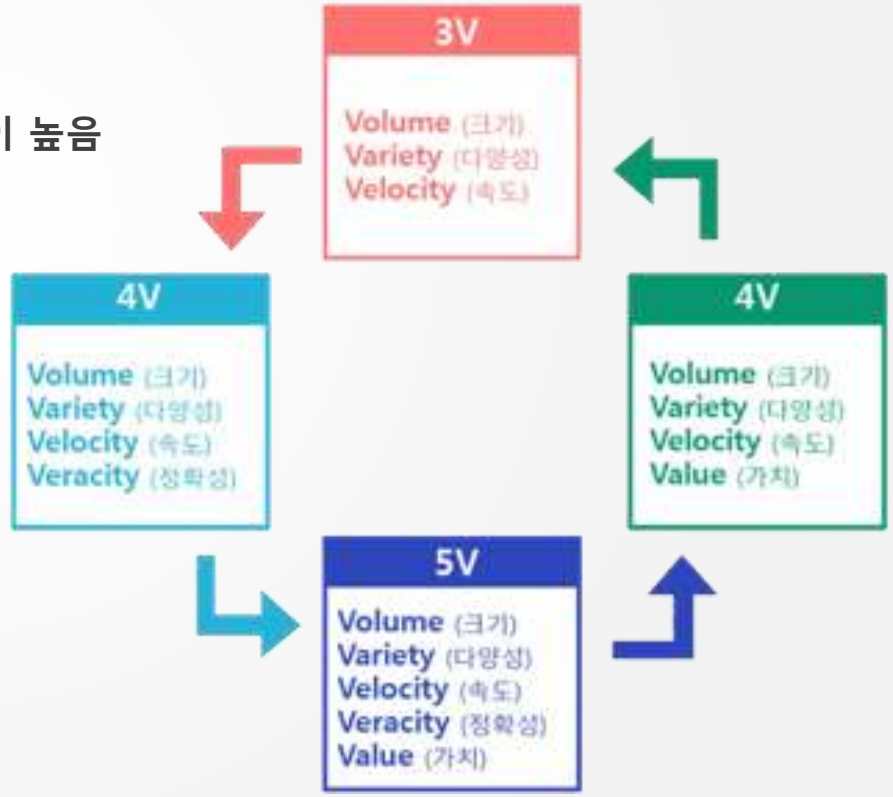
  데이터가 맥락에 따라 의미가 달라짐

- **시각화(Visualization)**

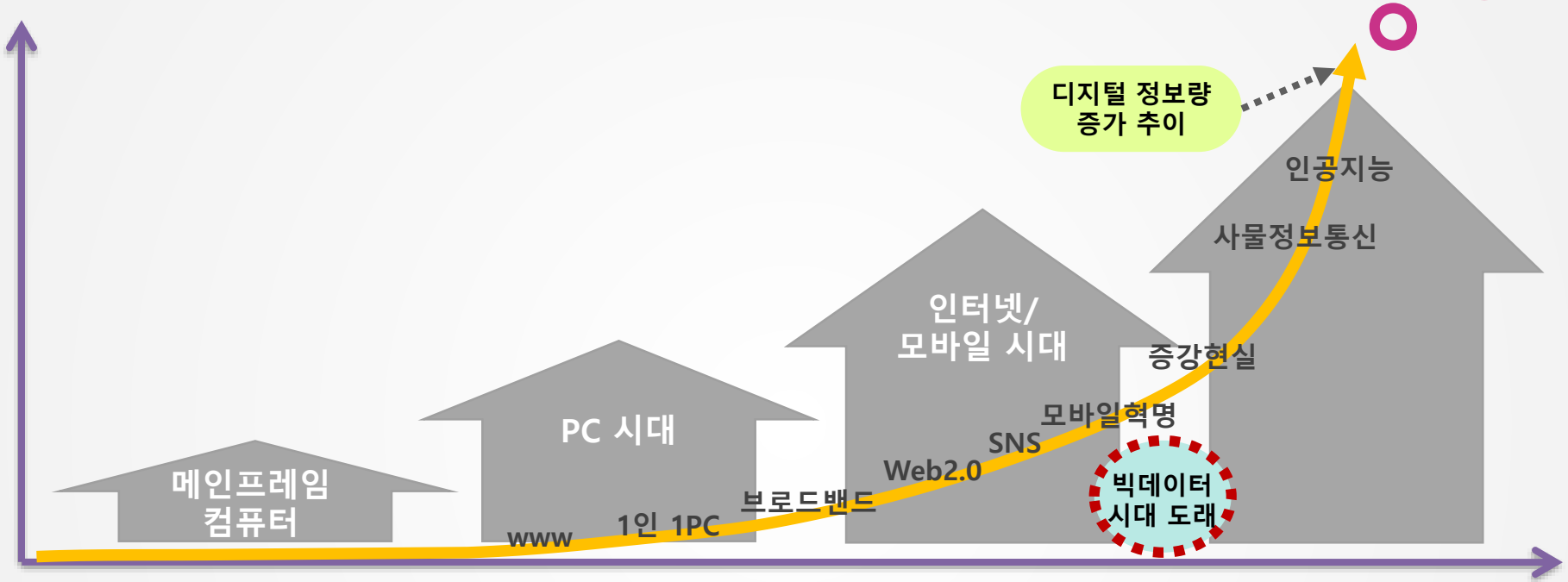
  정형 및 비정형 데이터를 수집하여

  복잡한 분석을 실행한 후 용도에

  맞게 가공하는 작업

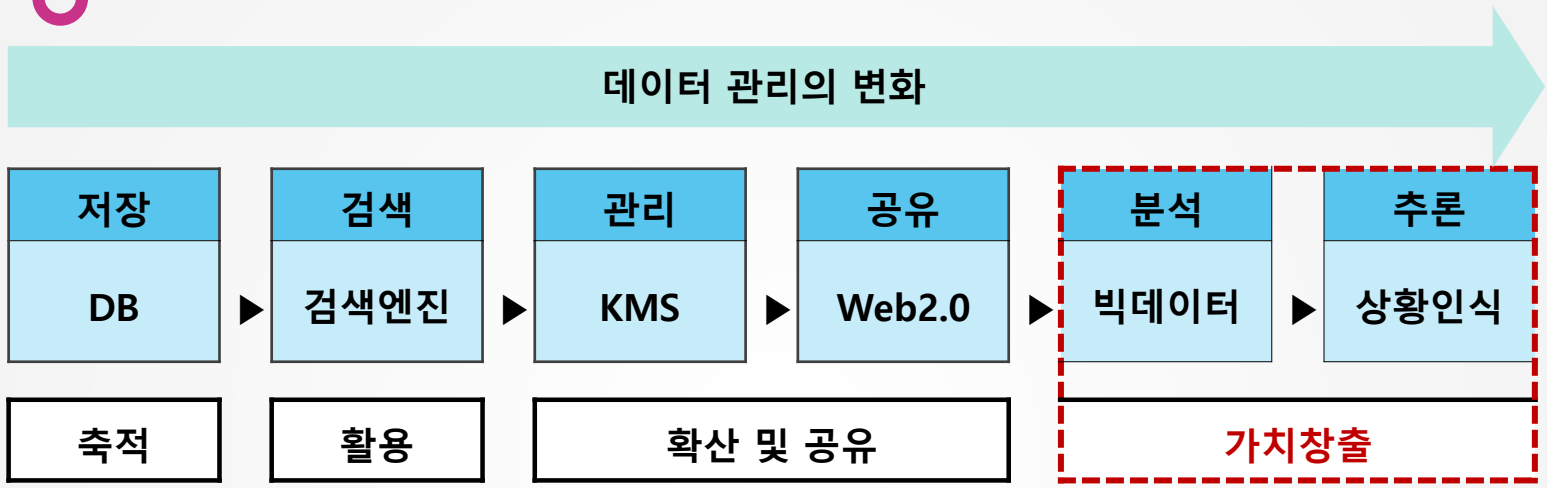


빅데이터 (Big Data) 증가



데이터 규모	EB(Exa Byte) (90년대 말=100EB)	ZB(Zetta Byte) 진입 (2011년=1.8ZB)	ZB 본격화 시대 ('20년='11년 대비 50배 증가)
데이터 유형	정형 데이터 (데이터베이스, 사무정보)	비정형 데이터 (이메일, 멀티미디어, SNS)	사물정보, 인지정보 (RFID, Sensor, 사물통신)
데이터 특성	구조화	다양성, 복합성, 소셜	현실성, 실시간성

[데이터 관리의 변화와 데이터 리터러시 역량]



데이터 리터러시 하위 역량

- 데이터 수집 : 필요한 데이터를 빠른 시간에 검색, 선별하여 확보하는 능력
- 데이터 관리 : 분석이 가능한 형태로 구조화, 정제하는 것
- 가공 및 분석 : 목적에 맞는 분석 방법을 사용해 의미 있는 결과 도출 능력
- 데이터 시각화 : 다른 사람이 이해할 수 있도록 그래프, 차트 등의 형태로 시각화
- 데이터 기획 : 전반적인 데이터 간의 관계를 이해하고 데이터 활용을 계획하는 능력

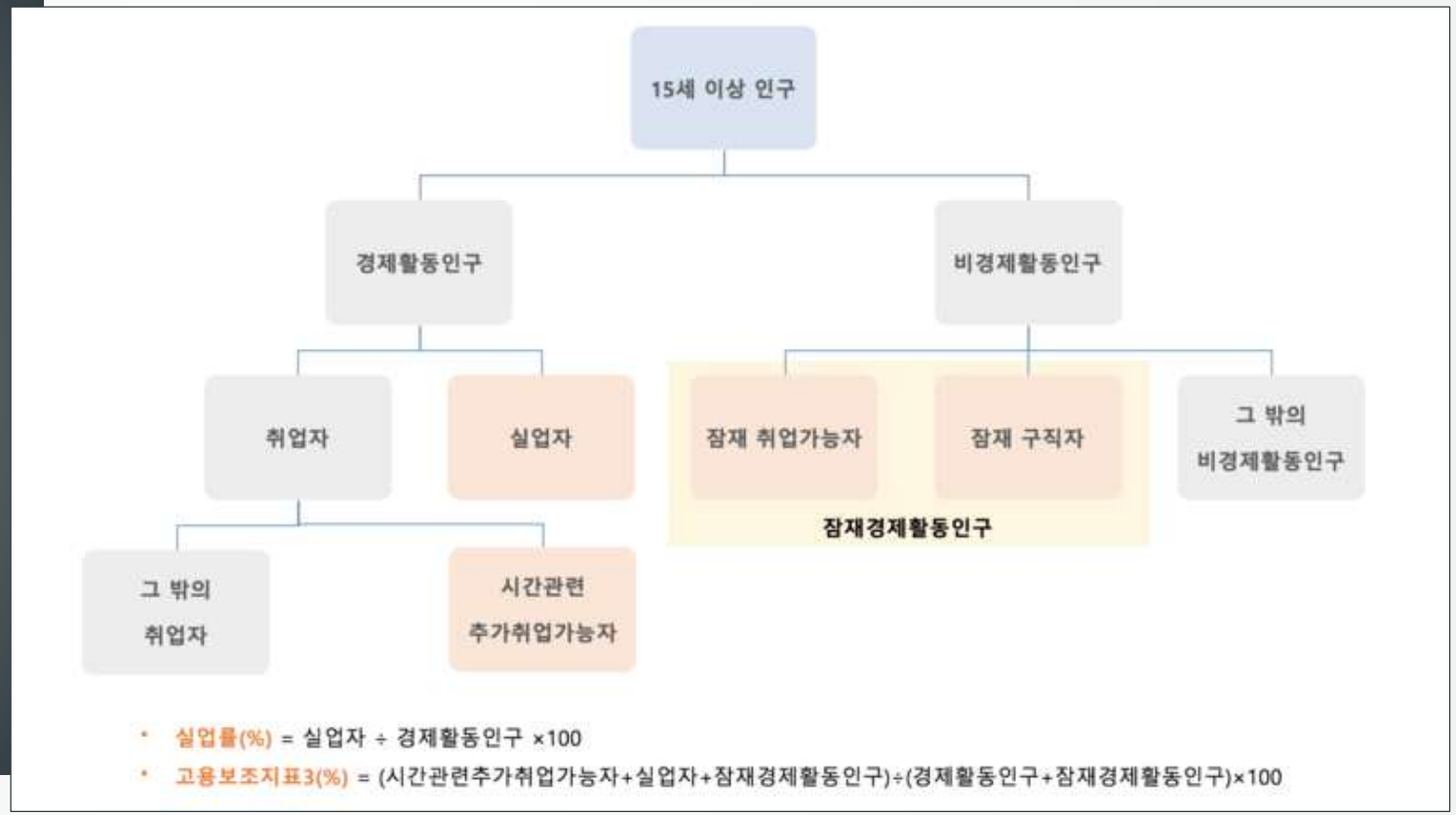
[일상 속 데이터 리터러시 : 통계 리터러시]

○ **통계 자료가 인용된 뉴스 기사나 보고서**

- 객관적이고 신뢰를 줄 수 있어 많이 사용함
- **데이터 해석시**

  측정배경, 방법, 기준 등을 고려해야 함

○ **기준에 따라 다른 지표**



- 통계청이 밝힌 2017년 12월 기준 우리나라의 실업률은 3.3%

  - 조사대상 주간에 수입 있는 일을 하지 않았고, 지난 4주간 일자리를 찾아 적극적으로 구직활동을 하였던 사람
- 고용보조지표3 (체감 실업률과 가장 가까운 지표)는 10.9%

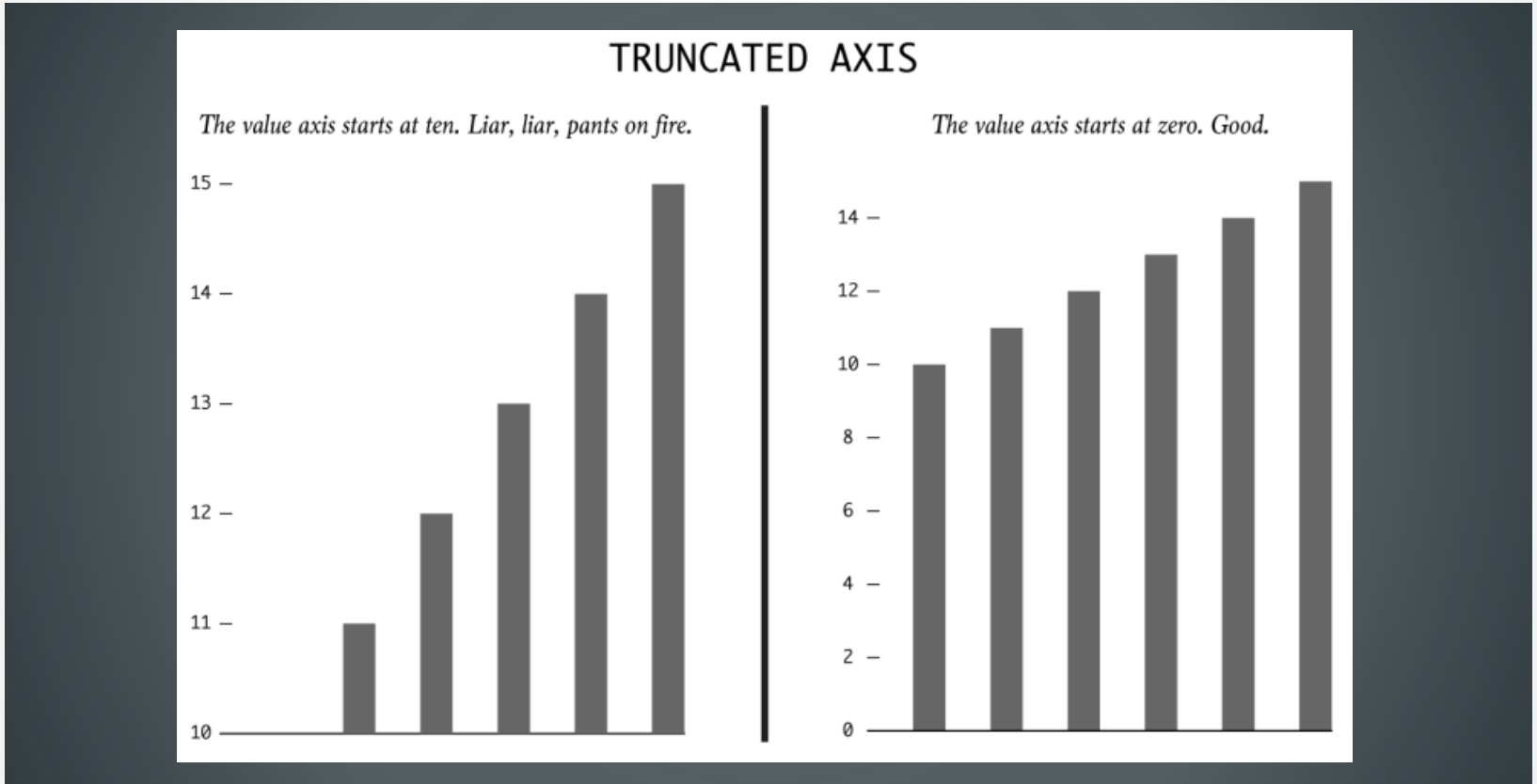
[일상 속 데이터 리터러시 : 시각화 리터러시]

- 시각화 자료 비판적으로 바라보기

  - 작성자는 자신의 의도에 따라 시각화 유형을 선택하고 활용

  - 장점이 될 수도 있지만 악용될 가능성도 있음

- 잘못된 시각화 차트 예 : 막대 그래프 축의 기준 값

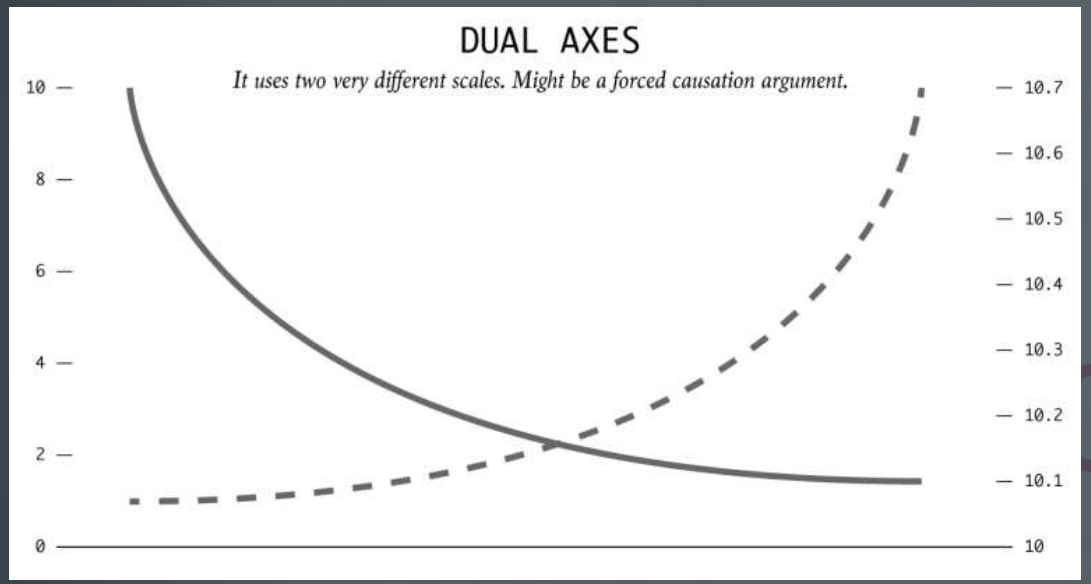


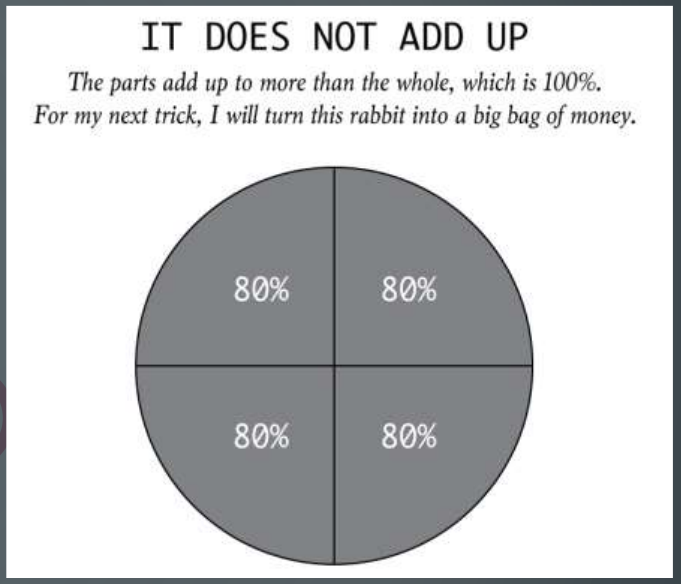
[일상 속 데이터 리터러시 : 시각화 리터러시]

잘못된 시각화 차트 예 : 이중축 사용

- 상관관계, 원인결과 분석 등에 사용
- 이중축(보조축) 사용으로 측정 기준별로 크기가 축소되거나 확장 가능

잘못된 시각화 차트 예 : 100%가 아닌 파이차트



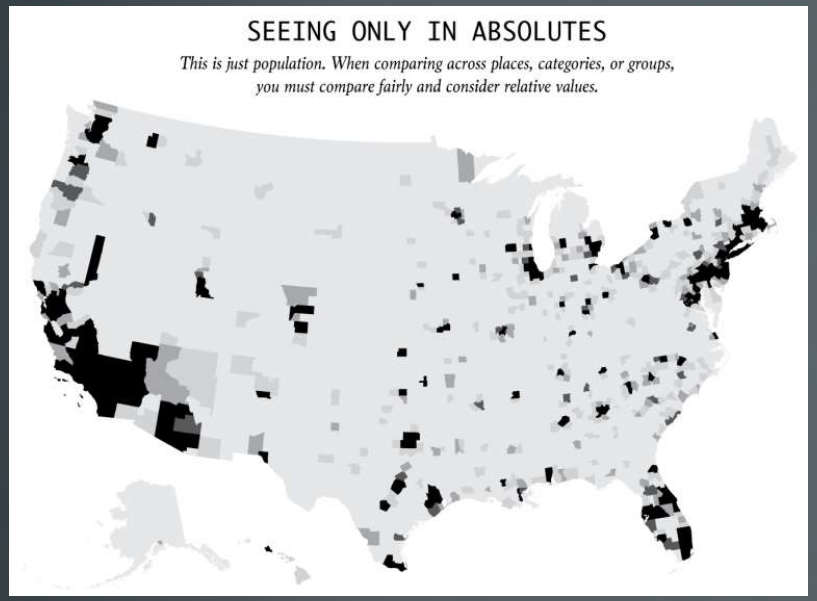


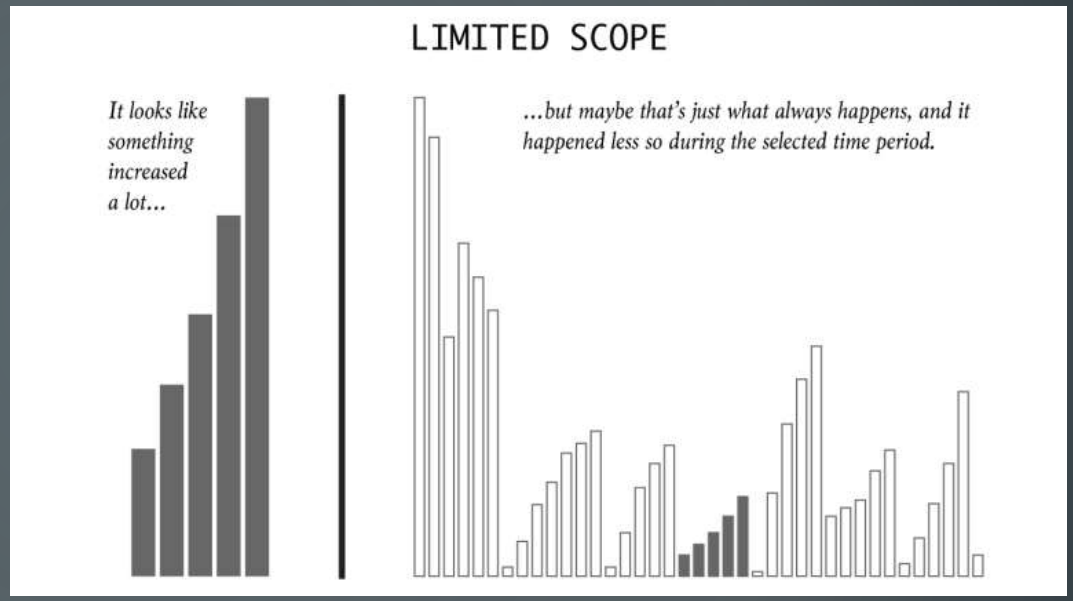
[일상 속 데이터 리터러시 : 시각화 리터러시]

- 잘못된 시각화 차트 예 : **절대값에 유의**

  - 첫번째 도시에는 10번의 강도사건이 있었고, 다른 도시에는 5번의 강도사건이 있었다.
    과연 첫번째 도시는 두번째 도시보다 위험한 도시인가?

- 잘못된 시각화 차트 예 : 날짜와 시간대를 임의적으로 변경하는 경우





[일상 속 데이터 리터러시 : 시각화 리터러시]

- 그래프 오류 예시

[잘못된 그래프 표현]

[수정된 그래프 표현]

THANK YOU
FOR LISTENING!